Disentangling unisensory from fusion effects in the attentional modulation of McGurk effects: a Bayesian modeling study suggests that fusion is attention-dependent

Jean-Luc Schwartz, Kaisa Tiippana, Tobias Andersen

Disentangling unisensory from fusion effects in the attentional modulation of McGurk effects: a Bayesian modeling study suggests that fusion is attention-dependent

Jean-Luc Schwartz[1], *Kaisa Tiippana*[2], *Tobias Andersen*[3]

[1] Gipsa-Lab, Speech & Cognition Department, CNRS-Grenoble University, France
[2] Institute of Behavioural Sciences, University of Helsinki, Finland
[3] Technical University of Denmark, Denmark

[1]Jean-Luc.Schwartz@gipsa-lab.grenoble-inp.fr, [2]kaisa.tiippana@helsinki.fi, [3]ta@imm.dtu.dk

Abstract

The McGurk effect has been shown to be modulated by attention. However, it remains unclear whether attentional effects are due to changes in unisensory processing or in the fusion mechanism. In this paper, we used published experimental data showing that distraction of visual attention weakens the McGurk effect, to fit either the Fuzzy Logical Model of Perception (FLMP) in which the fusion mechanism is fixed, or a variant of it in which the fusion mechanism could be varied depending on attention. The latter model was associated with a larger likelihood when assessed with a Bayesian Model Selection criterion. Our findings suggest that distraction of visual attention affects fusion by decreasing the weight of the visual input.

Index Terms: McGurk effect, attention, FLMP, modeling

1. Introduction

While it had been initially claimed since McGurk and MacDonald (1976) [1] that the McGurk effect (conflicting visual speech altering the auditory speech percept due to audio-visual fusion) was automatic and not under the control of attention, it appeared later that instruction to attend more to audition or to vision might bias perception [2]. More recently, Tiippana et al. (2004) [3] showed that if attention is distracted from visual speech by the presentation of a concurrent visual stimulus (a leaf superimposed on the speaking face), the role of visual speech decreases in fusion, so that the McGurk effect gets weaker. The authors modeled their data with the Fuzzy Logical Model of Perception (FLMP) [2], which provided a good fit as assessed with the root mean square error (RMSE). Since the FLMP entails a fixed integration rule, a good fit of the model suggests that the attentional effect acts on the visual input, rather than on fusion. However, they also noted that in the experimental data there was little evidence of unisensory visual attentional effects. This discrepancy is possible because of the non-linearity of the FLMP, which allows small, statistically non-significant differences in visual response probabilities to cause large, significant changes in audiovisual response probabilities. Tiippana et al. concluded that this discrepancy prevented them from being able to determine whether the attentional manipulation influenced unisensory processing or fusion based on FLMP fits.

Schwartz (2006) [4] argued that the good fits of the FLMP to McGurk data might be due to over-fitting. He showed that the error function of the FLMP has a very steep slope in that area of parameter space, which models the McGurk illusion. This means that a small change in the parameters can cause a large change in the model likelihood. The model is, in other words, very flexible in that its parameters can be nudged to accommodate almost any data set, particularly for conflicting auditory and visual inputs, as in the McGurk effect. This is the hallmark of over-fitting. The problem with over-fitting is that, although the model fits well, it generalizes poorly: The model with parameters fit to one data set does a poor job in describing another, very similar dataset. In order to overcome this difficulty, one needs to take the entire likelihood function into account rather than just its maximum. This is the principle of the Bayesian Model Selection (*BMS*) criterion. This criterion involves computation of the global likelihood of a model considering a set of experimental data, which is computationally complex, but Schwartz introduced the so-called Laplace approximation (*BMSL*), which appears to be easy to implement and compute.

In a later study, Schwartz [5] introduced a variation of the FLMP, the weighted FLMP (WFLMP), in which inputs from audition and vision are weighted. He compared the two models using various criteria: the *RMSE*, the *RMSE* corrected for the number of free parameters and the *BMSL*. He found that all measures favored the WFLMP. Closer inspection revealed that the *RMSE* based measures always favored the model with more free parameters, which could be due to over-fitting. The *BMS* did not show this behavior indicating that it is less influenced by over-fitting.

In addition to good fits and ability to generalize a good model should also add to our qualitative understanding of the underlying cognitive processes. Schwartz showed that the WFLMP did that since its weights provided a meaningful indicator of how much individual observers relied on audition versus vision. This issue is similar to Tiippana's question whether an irrelevant visual object can distract visual attention and thereby decrease the weight of visual information in audiovisual speech perception. Therefore, Schwartz' approach might help resolve the paradox that Tiippana et al. encountered. Hence, in the current study, we examine this issue by comparing data fits provided by various implementations of the WFLMP vs. the normal FLMP, by minimizing the *BMSL* criterion to Tiippana et al.'s data.

2. Methods

2.1. Experimental data

In Tiippana & al. (2004) [3], subjects (n=14) recognized consonants /k/, /p/ and /t/ in /eCe/ context, presented in extended factorial design in two conditions: attend Face and attend Leaf. In the latter, subjects attended to a leaf floating across the talker's face instead of the face. The data consisted of response distributions to 15 stimuli (3 auditory A, 3 visual V, 9 audiovisual AV where 3 were congruent A=V and 6 incongruent A≠V i.e. McGurk stimuli) in 2 conditions (Face

and Leaf) for 5 response categories: /k/, /p/, /t/, combination (combination of A and V consonants) and 'other' (Fig. 1).

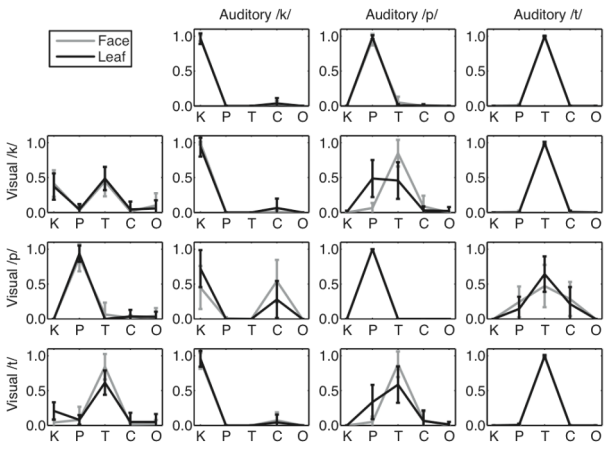

Figure 1 – *Summarized results of the experimental data in [3]. Each plot provides mean response proportions for the 5 response categories: /k/, /p/, /t/, combination (C) and 'other' (O). Error bars indicate the standard error of the mean across subjects. Plots are arranged so that rows indicate the auditory stimulus (none, /k/, /p/ or /t/) and columns indicate the visual stimulus (none, /k/, /p/ or /t/). Grey lines denote the attend Face condition and black lines the attend Leaf condition*

2.2. Models

Two models were fitted to the data. Firstly, the FLMP:

$$P(R_i \mid A,V) = \frac{P(R_i \mid A)P(R_i \mid V)}{\sum_j P(R_j \mid A)P(R_j \mid V)}$$

where R_i and R_j are response categories, A and V are auditory and visual stimuli, $P(R_i|A)$, $P(R_i|V)$ and $P(R_i|A,V)$ are auditory, visual and audiovisual response probabilities, respectively. Considering the Face and Leaf conditions, the equations are:

$$P(R_i \mid A_{face},V_{face}) = \frac{P(R_i \mid A_{face})P(R_i \mid V_{face})}{\sum_j P(R_j \mid A_{face})P(R_j \mid V_{face})}$$

$$P(R_i \mid A_{leaf},V_{leaf}) = \frac{P(R_i \mid A_{leaf})P(R_i \mid V_{leaf})}{\sum_j P(R_j \mid A_{leaf})P(R_j \mid V_{leaf})}$$

The second model was the WFLMP:

$$P(R_i \mid A,V) = \frac{P(R_i \mid A)^{\lambda_A} P(R_i \mid V)^{\lambda_V}}{\sum_j P(R_j \mid A)^{\lambda_A} P(R_j \mid V)^{\lambda_V}}$$

where λ_A and λ_V are factors used to weight the auditory and visual inputs in the computation of the audiovisual responses (see other introductions of weights inside the FLMP in [6]). For each condition (Face or Leaf), we define a λ value between 0 and 1, and compute λ_A and λ_V from λ by: $\lambda_A = \lambda / (1 - \lambda)$ and $\lambda_V = (1 - \lambda) / \lambda$. Therefore the weighted model WFLMP needs two more parameters than FLMP (λ_{Face} and λ_{Leaf}).

2.3. Criteria for models assessment

The assessment criteria applied were *RMSE*, *corrected RMSE* and *BMSL*, which all give smaller values, the better the model fit. Let us consider a speech perception experiment for categorization of speech stimuli involving n_E experimental conditions E_j, and in each condition, n_C possible responses corresponding to different phonetic categories C_i. In most papers comparing models in the field of audiovisual speech perception, the tool used to compare models is the fit estimated by the root mean square error *RMSE*, computed by taking the squared distances between observed and predicted probabilities of responses, averaging them over all categories C_i and all experimental conditions E_j, and taking the square root of the result:

$$RMSE = \sqrt{\sum_{i,j} \left[P_j(R_i \mid A,V) - p_j(R_i \mid A,V)\right]^2 \Big/ n_E n_C}$$

where observed probabilities are in lower case and model probabilities in upper case.

Considering that two models might differ in their number of degrees of freedom, Massaro (1998) proposes to apply a correction factor *k/(k-f)* to *RMSE*, where k is the number of data points and f the number of degrees of freedom of the model. This provides the second criterion, the *corrected RMSE*:

$$RMSE_{cor} = \frac{k}{(k-f)} RMSE$$

The third criterion used here is the Laplacian approximation to the Bayesian Model Selection (BMSL) criterion. If $\mathbf{D}$ is a set of k data points d_i, and M a model with parameters $\boldsymbol{\Theta}$, the best fit is the maximum of the likelihood of the model given the data set, that is the value of $\boldsymbol{\Theta}$ maximizing $L(\boldsymbol{\Theta}|M) = P(\mathbf{D}|\boldsymbol{\Theta},M)$. However, the maximally likely parameter set is not the only possible parameter set. By assessing models by comparing only their maximum likelihoods we ignore the possibility that this is not the true underlying model. BMS estimates the likelihood integrated over all parameter values [7]:

$$BMS = -\log \int L(\Theta \mid M)P(\Theta \mid M)d\Theta$$

Bayesian Model Selection has already been applied to the comparison of AV speech perception models, including FLMP [8, 9].

The computation of BMS through this equation is complex. It involves the estimation of an integral, which generally requires use of numerical integration techniques, typically Monte-Carlo methods. However, Jaynes (1995, ch. 24, [10]) proposes an approximation of the total likelihood based on an expansion of log(L) around the maximum likelihood point $\boldsymbol{\theta}$:

$$\log(L(\Theta)) \approx \log(L(\theta)) + \frac{1}{2}(\Theta - \theta)^T \left[\partial^2 \log(L)/\partial\Theta^2\right]_\theta (\Theta - \theta)$$

where $[\partial^2 log(L) / \partial\boldsymbol{\Theta}^2]_{\boldsymbol{\theta}}$ is the Hessian matrix of the function *log(L)* computed at the position of the parameter set $\boldsymbol{\theta}$ providing the maximal likelihood L_{max} of the considered

model. This leads to the so-called Laplace approximation of the BMS criterion [11]:

$$BMSL = -\log(L_{\max}) - \frac{m}{2}\log(2\pi) + \log(V) - \frac{1}{2}\log(|\Sigma|)$$

where V is the total volume of the space occupied by parameters $\boldsymbol{\Theta}$, m is its dimension, that is the number of free parameters in the considered model, and $\boldsymbol{\Sigma}$ is defined by:

$$\Sigma^{-1} = \left[\partial^2 \log(L)/\partial\Theta^2\right]_\theta$$

The preferred model considering the data **D** should *minimize* the *BMSL* criterion. There are in fact three kinds of terms in the computation of *BMSL*. Firstly, the term $-\log(L_{max})$ is directly linked to the maximum likelihood of the model, more or less accurately estimated by *RMSE*: the larger the maximum likelihood, the smaller the *BMSL* criterion. Then, the two following terms are linked to the dimensionality and volume of the considered model. Altogether, they result in the handicapping of models that are too "large" (that is, models with a too high number of free parameters). Finally, the fourth term provides a term favoring models with a large value of $\det(\boldsymbol{\Sigma})$. Indeed, if $\det(\boldsymbol{\Sigma})$ is large, this means that the determinant of the Hessian matrix of $\log(L)$ is small, which expresses that the likelihood L does not vary too quickly around its maximum value L_{max}.

BMSL has the double interest to be easy to compute, and easy to interpret in terms of fit and stability. Furthermore, if the amount of available data is much greater than the number of parameters involved in the models to compare (that is, the dimension m of the $\boldsymbol{\Theta}$ space) the probability distributions become highly peaked around their maxima, and the central limit theorem shows that the approximation of *BMS* by *BMSL* becomes quite reasonable. Kass & Raftery (1995) [11] suggest that the approximation should work well for a sample size greater than 20 times the parameter size m. In our case, the ratio will be from 24 (in the model with the highest number of parameters, that is 50) to 100 (for the model with 12 parameters) and even 600 (for the smallest model with 2 parameters).

2.4. Varying the number of free parameters

The number of free parameters in most model comparison studies in AV speech perception is generally kept fixed to the "natural number of degrees of freedom" of the model, that is the number of free parameters necessary to implement the model in its most extensive definition. Care is generally taken to check that the models have basically the same number of degrees of freedom, otherwise the *RMSE* correction previously described could be applied. Notice that this correction looses some sense if a parameter is introduced with no effect on the model likelihood (a "useless parameter") while *BMSL* naturally discards useless parameters.

Of course, completely useless parameters generally do not exist, since this would correspond to some kind of misconception of the model. However, it is important to assess the possibility that some parameters are not really useful in the model behavior. For example, while all model comparisons generally involve a subject-by-subject assessment – and it will also be the case here – it may be interesting to test if some parameters could not in fact be similar from one subject to the other. The same could be done from one experimental condition to the other. Therefore, we systematically tested various implementations of the models to compare, with a progressively increasing number of fixed parameters and thus a decreasing number of free parameters, in order to attempt to determine the *true* number of degrees of freedom of the model, that is the number of free parameters really useful, and providing the highest global likelihood of the model knowing the data. Our basic assumption is that it is under the condition of true number of degree of freedom that models can be really assessed and compared in sound conditions.

The logic guiding these progressive constraints decreasing the number of free parameters is that if the FLMP exploits the available free parameters to adapt its behavior to any pattern of experimental data (see [4]), it will have both a very low *RMSE* (even corrected) and a high *BMSL* (poor global likelihood), and hence it is necessary to take out as many free parameters as possible to really assess the regularity and consistency of subjects' behavior.

We compared six variants of the FLMP and WFLMP models. The baseline model is the 48-parameter FLMP_48, fitting 48 values for each subject, i.e. values of $P(R_i/A_{Face})$, $P(R_i/V_{Face})$, $P(R_i/A_{Leaf})$ and $P(R_i/V_{Leaf})$ for the three auditory and the three visual stimuli and 4 responses (/k/, /p/, /t/, "comb", the fifth one "other" being provided by normalization to 1). The corresponding WFLMP_50 model uses 50 parameters per subject, that is, the 48 previous ones plus λ_{Face} and λ_{Leaf}.

Then we tested five variants involving various decreases of the number of free parameters:

In FLMP_36, we assumed that since visual but not auditory attention was manipulated in the experiments, there is no difference between Face and Leaf conditions for auditory-only responses, so that all values of $P(R_i/A_{Face})$ and $P(R_i/A_{Leaf})$ are equal, fixing 12 free parameters. WFLMP_38 is the same plus λ_{Face} and λ_{Leaf}.

In FLMP_24, we further assumed that responses to the auditory stimuli do not differ between subjects due to near-perfect recognition (96-100% correct), so that $P(R_i/A_{Face})$ and $P(R_i/A_{Leaf})$ are both equal to each other and the same for all subjects. This was done through a Round Robin technique, in which a given parameter for one subject is estimated from the mean value taken by the parameter in the whole corpus excluding the current subject from the computation. This technique, classical and computationally simple, separates the data used to estimate the parameter from the data used to test the model. The parameter is therefore not free because it is not adjusted to accommodate the test data. Instead, it is fixed by independent data. This reduced the number of free parameters by 12 leaving 12 free parameters per subject for $P(R_i/V_{Face})$ and $P(R_i/V_{Leaf})$ each. WFLMP_26 is the same plus λ_{Face} and λ_{Leaf}.

In FLMP_12, we further assumed that visual responses are the same in the Face and Leaf conditions, fixing 12 more parameters. WFLMP_14 is the same plus λ_{Face} and λ_{Leaf}.

FLMP_16 is a variant of FLMP_12 in which we added four free parameters enabling responses to visual /t/ differ between Face and Leaf conditions since this was the only statistically significant difference for visual-only stimuli in the experimental data. WFLMP_18 is the same plus λ_{Face} and λ_{Leaf}.

Finally, WFLMP_2 is a variant of WFLMP in which we assumed that all subjects have the same auditory-only and visual-only responses. That is, $P(R_i/A_{Face})$ is equal to $P(R_i/A_{Leaf})$, and $P(R_i/V_{Face})$ is equal to $P(R_i/V_{Leaf})$ and their values are identical for all subjects. The only free parameters are now λ_{Face} and λ_{Leaf}. Of course, there is no FLMP counterpart, since there would be no free parameter.

3. Results

The models were first assessed in terms of the *RMSE* (Fig. 2), which decreases as the number of parameters increases, as can be seen in Fig. 2 for both FLMP and WFLMP. However, the *RMSE* is lowest and almost the same for the two model variants with the highest number of parameters: FLMP_48/WFLMP_50 and FLMP_36/WFLMP_38. The latter variant is almost as good as the former despite the large decrease in parameters since in the data there is almost no difference in auditory-only responses between Face and Leaf conditions. However, the difference between FLMP and WFLMP becomes quite large for FLMP_16/WFLMP_18 and other variants with fewer parameters, so that WFLMP variants provide smaller *RMSE* values than the corresponding FLMP variants due to the two additional parameters.

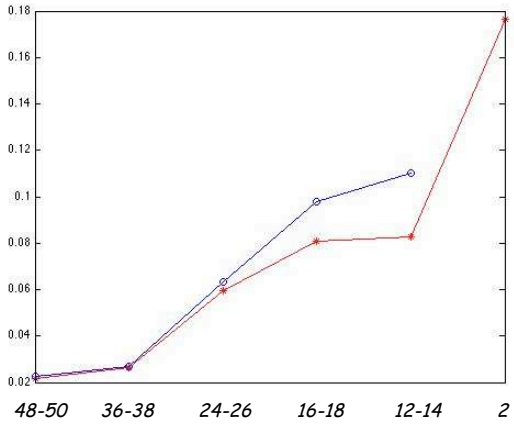

Figure 2 – RMSE for the FLMP (in blue) and WFLMP (in red) for the corresponding degrees of freedom (48-50 for FLPM-48 and WFLMP-50, etc; 2 for WFLMP-2)

Correcting for the degrees of freedom does not change the pattern much, as can be seen by comparing Fig. 3 showing the corrected *RMSE* with Fig. 2. Again, there are four equally good models, FLMP_48, FLMP_36, WFLMP_50 and WFLMP_38.

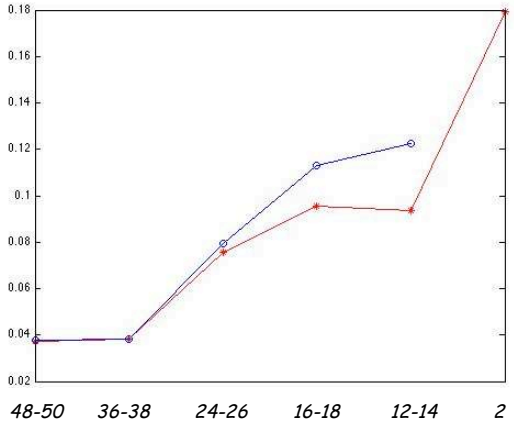

Figure 3 – Corrected RMSE for FLMP (blue) and WFLMP (red). Other details as in Fig. 2.

When assessing the models using the *BMSL*, the pattern is very different (Fig. 4). The *BMSL* decreases as the number of parameters decreases for all paired FLMP/WFLMP variants. It might seem puzzling that FLMP_12/WFLMP_14 have the lowest BMSL since here it is assumed that there is no difference between Face and Leaf conditions for visual responses, even though Tiippana et al. showed that there was a statistically significant difference for visual /t/. Probably these differences are of a second order compared with the basic phenomenon captured by the FLMP that audiovisual responses are well modeled by a multiplicative process.

The main finding here is that the best FLMP variant with 12 parameters is significantly poorer than the best WFLMP variant with 14 parameters, as shown by a Wilcoxon signed rank test. Finally, the rise of the *BMSL* curve for WFLMP_2 shows that *BMSL* is not just driven by the trend to decrease with the number of degrees of freedom: there is indeed a minimum, here for 14 degrees of freedom, and for the WFLMP rather than the FLMP. In the same vein, adding two free parameters from FLMP to WFLMP can lead to either a *BMSL* increase (for the first variants with too many free parameters) or a *BMSL* decrease for the last variants with fewer parameters.

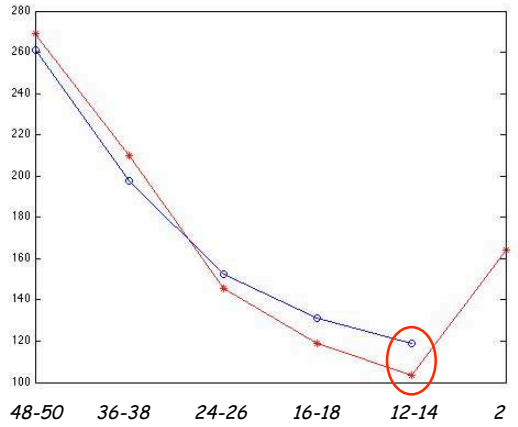

Figure 4 – BMSL for FLMP (blue) and WFLMP (red). The red ellipse marks the best variants. Other details as in Fig. 2. The red ellipsis marks the best configurations for FLMP and WFLMP

4. Discussion

This modeling work enables us to really assess the informational content of the provided experimental material in sound terms, relating unisensory and multisensory data in a coherent way thanks to Bayesian Model Selection. It appears that the WFLMP is associated with a larger global likelihood than FLMP, which suggests that in these data, there is indeed a modulation of fusion by attentional distraction, inducing a decrease in fusion per se, possibly superimposed with a modification in unisensory visual performance. This adds to the growing literature on attentional modulation of fusion in the McGurk effect (e.g. [12] and also the paper submitted by Nahorna et al., in the present conference).

This work extends methodological developments by the authors [4, 5, 13, 14] by confirming the superiority of the BMS approach to the *RMSE* approach in model assessment. However, while the *BMS* seems to provide an efficient criterion for model assessment and comparison, other tools could be used in future experiments, including cross-validation which provides a functional way to assess stability of the best fit, probably coherent with the *BMS*. But most importantly, the *BMS* (with its *BMSL* easy-to-compute approximation) together with the variable-degrees-of-freedom technique, should be

used to re-assess various audiovisual fusion experiments on e.g. attentional [15], developmental [16] or cross-linguistic [17, 18] effects on audiovisual in speech perception.

5. Acknowledgements

We thank Prof. Mikko Sams for his collaboration in ref. [3] and other related previous work.

6. References

[1] McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746-748.

[2] Massaro, D.W. (1998). *Perceiving Talking Faces*. Cambridge: MIT Press.

[3] Tiippana, K., Andersen, T.S., & Sams, M., (2004). Visual attention modulates audiovisual speech perception. *European Journal of Cognitive Psychology*, *16*, 457-472.

[4] Schwartz, J.L. (2006). Bayesian model selection: The 0/0 problem in the Fuzzy-Logical Model of Perception. *J. Acoust. Soc. Am., 120*, 1795-1798.

[5] Schwartz, J.L. (2010). A reanalysis of McGurk data suggests that audiovisual fusion in speech perception is subject-dependent. *J. Acoust. Soc. Am., 127*, 1584-1594.

[6] Schwarzer, G., & Massaro, D.W. (2001). Modeling face identificaton processing in children and adults. *Journal of Experimental Child Psychology, 79*, 139-161.

[7] Pitt, M.A., & Myung, I.J. (2002). When a good fit can be bad. *Trends in Cognitive Science, 6*, 421-425.

[8] Massaro, D.W., Cohen, M. M., Campbell, C.S., & Rodriguez, T. (2001). Bayes factor of model selection validates FLMP. *Psychonomic Bulletin & Review, 8*, 1-17.

[9] Pitt, M.A., Kim, W., & Myung, I.J. (2003). Flexibility versus generalizablity in model selection. *Psychonomic Bulletin & Review, 10*, 29-44.

[10] Jaynes, E. T. 1995. *Probability Theory: The logic of Science*. Cambridge : Cambridge University Press.

[11] Kass, R.E., & Raftery, A.E. (1995). Bayes factor. *Journal of the American Statistical Association, 90,* 773-795.

[12] Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Curr Biol., 15*, 839-843.

[13] Andersen, T. S., Tiippana, K., Lampinen, J., & Sams, M., 2001. Modeling of audiovisual speech perception in noise. *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP 2001)*, pages 172-176.

[14] Andersen, T. S., Tiippana, K., & Sams, M., 2002. Using the fuzzy logical model of perception in measuring integration of audiovisual speech in humans. *Proceedings of NeuroFuzzy2002*.

[15] Tuomainen, J, Andersen, T.S., Tiippana, K, & Sams, M. (2005). Audio-visual speech perception is special. *Cognition, 96*, B13–B22.

[16] Sekiyama. K., & Burnham, D. (2004). Issues in the development of auditory-visual speech perception: Adults, infants and children. In *Proceedings of the 8th International Conference on Spoken Language Processing*, edited by Soon Hyob Kim & Dae Hee Yuon (Seoul, Sunjin Printing, Korea), pp 1137-40.

[17] Sekiyama, K., & Tokhura, Y. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *J. Acoust. Soc. Am., 90*, 1797-1825.

[18] Massaro, D.W., Cohen, M.M., Gesi, A., Heredia, R., & Tsuzaki, M. (1993). Bimodal speech perception: An examination across languages. *J. Phonetics, 21*, 445-478.